DOCUMENT RESUME

ED 450 129 TM 032 319

AUTHOR Veldkamp, Bernard P.
TITLE Constrained Multidimensional Test Assembly. Research Report.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO RR-00-04
PUB DATE 2000-00-00
NOTE 30p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Item Banks; *Test Construction; Test Items
IDENTIFIERS Constraints; *Mathematical Programming; *Multidimensionality (Tests)

ABSTRACT

Two mathematical programming approaches are presented for the assembly of ability test from item pools calibrated under a multidimensional item response theory model. Item selection is based on Fisher's Information matrix. Several criteria can be used to optimize this matrix. In this paper, the A-criterion and the D-criterion are applied. In a mathematical programming approach, both criteria provide good results for the two dimensional case. Empirical examples for a two-dimensional mathematics item pool illustrate the methods. Recommendations are provided about when to apply either approach. (Contains 1 table, 2 figures, and 23 references.) (Author/SLD)

# Constrained Multidimensional Test Assembly

Bernard P. Veldkamp

TM

# Constrained Multidimensional Test Assembly

**Research Report 00-04**

Bernard P. Veldkamp



1







*faculty of*
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

**Abstract**

Two mathematical programming approaches are presented for the assembly of ability tests from item pools calibrated under a multidimensional IRT model. Items selection is based on Fisher's Information matrix. Several criteria can be used to optimize this matrix. In this paper the A-criterion and the D-criterion are applied. In a mathematical programming approach, both criteria provide good results for the two dimensional case. Empirical examples for a two-dimensional mathematics item pool illustrate the methods. Recommendations are provide about when to apply either approaches.

Keywords: Greedy heuristic, linear approximation, mathematical programming, multidimensional IRT, optimal test assembly, test design.

## Introduction

In educational measurement, item response theory (IRT), (Birnbaum,1968), is generally used as a psychometric theory to govern the test assembly process. In this process, three steps can be distinguished. First an IRT model is chosen and the items in the item bank are calibrated. From an item bank many different tests can be assembled. Therefore, the second step consists of specifying the properties of the desired test, for example, the test length, the desired amount of information in the test, or the administration time needed for the test. The third step of the process is to choose an algorithm that selects items from the item bank such that the test specifications are met. A mathematical programming approach is often used for this step.

The idea of using a mathematical programming approach was suggested by Yen (1983) and Theunissen (1985). Ever since, several papers have proposed various LP algorithms and heuristics to solve test assembly problems. Recently, van der Linden (1996), Segall (1996) and Luecht (1996) addressed the subject of assembling tests measuring multiple abilities, i.e. from an item bank calibrated using a multidimensional IRT (MIRT) model.

Measuring multiple abilities using a MIRT model is not always seen as a practical option (Wainer et al. 1990). However, the main advantage of MIRT is increased measurement efficiency (Segall, 1996). When the dimensions measured in the test have non-zero correlations, items with a content classification in one dimension provide information about the other dimensions. MIRT models have been subject of research for many years (Bock, Gibbons and Muraki, 1988, Ackerman, 1994; McKinley and Reckase, 1983, and Reckase, 1985). Much research has been carried out in order to decide whether correlations are low enough to represent significant dimensions (McDonald, 1981; Reckase, 1979; Stout, 1987). Once a significant number of dimensions has been confirmed, the item parameters can be estimated. Programs as NOHARM (Fraser and McDonald, 1988) and TESTFACT (Wilson, Wood, and Gibbons, 1987) are often used in

this step. Based on these parameters, items can be selected from an item bank, to assemble specified tests.

In van der Linden (1996), an algorithm to solve the problem of assembling constrained tests that measure multiple traits was provided. However, intervention by the test assembler was needed to find the best solution. Therefore, the purpose of the present study was to find a heuristic that provides good solutions to the problem of assembling tests measuring multiple traits in a fully automated fashion. Two algorithms were developed. The first algorithm is based on a linear approximation of the objective function. Second, Luecht's (1996) algorithm for adaptive testing is adjusted in order to apply it to the problem of assembling constrained multidimensional P&P tests.

In the remainder of the paper, the MIRT model used for calibrating the item is described. Then a multidimensional version of a maximin model (van der Linden & Boekkooi-Timminga, 1989) is presented. Subsequently, two algorithms developed to assemble optimal tests are introduced. Both algorithms are compared by applying them to empirical examples. Finally both algorithms are discussed and recommendations for their use are provided.

## A Linear Logistic MIRT Model

### The MIRT Model

The model considered in this paper is a generalization of the two-parameter logistic model (Lord, 1980) to the multidimensional case (Reckase,1985). It can be formulated in the following manner:

$$P_i(\boldsymbol{\theta}_j) \equiv P(U_{ij} = 1|(\mathbf{a}_i, d_i, \boldsymbol{\theta}_j)) \quad (1)$$

$$\equiv \frac{e^{(\mathbf{a}_i \cdot \boldsymbol{\theta}_j + d_i)}}{1 + e^{(\mathbf{a}_i \cdot \boldsymbol{\theta}_j + d_i)}}, \quad (2)$$

where $P_i(\boldsymbol{\theta}_j)$ is the probability that a person $j = 1 \ldots J$ with ability vector $\boldsymbol{\theta}_j$ gives a correct response $U_{ij}$ to an item $i = 1 \ldots I$, $\mathbf{a}_i$ is the vector of discrimination parameters

of item $i$ along the abilities $\theta_{j1} \ldots \theta_{jm}$, $m$ is the dimensionality of the ability space, and $d_i$ is the parameter representing the difficulty of the item. In this paper, the item parameters are supposed to be known and the model is used to estimate the ability vectors $\boldsymbol{\theta}_j$ from realizations of the response variables $U_{ij} = u_{ij}$ for $i = 1 \ldots I$ and $j = 1 \ldots J$.

**Fisher's Information**

In the multiparameter case, Fisher's information is a matrix instead of a scalar (Lehmann, 1983; Segall 1996). Therefore, the (asymptotic) variances of the MLEs of the ability parameters $\theta_1, .., \theta_m$ are given by the diagonal elements of the inverse of Fisher's information matrix. For notational simplicity we consider the case of $m = 2$. Then Fisher 's information matrix, and the variance-covariance matrix are of the following forms:

$$I(\boldsymbol{\theta}) = \begin{bmatrix} \sum_{i=1}^{n} a_{1i}^2 P_i Q_i & \sum_{i=1}^{n} a_{1i} a_{2i} P_i Q_i \\ \sum_{i=1}^{n} a_{1i} a_{2i} P_i Q_i & \sum_{i=1}^{n} a_{2i}^2 P_i Q_i \end{bmatrix} \tag{3}$$

and

$$V(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = I(\boldsymbol{\theta})^{-1}. \tag{4}$$

In order to optimize measurement precision, either Fisher 's Information matrix has to be maximized or the variance-covariance matrix has to be minimized.

## A Multidimensional Maximin Model

A maximin model was used to formulate the test assembly process. The model consists of an objective function that has to be optimized over the set of possible tests meeting the specifications. The set is typically delineated by a number of mathematical constraints. But before the model is formulated, decision variables are introduced.

In the above mentioned matrices, sums are taken over the $n$ items in the test. In the test assembly process, it is unknown which items will be choosen in the test beforehand. To guarantee that the matrices are calculated for the items that are in the test, decision variables $x_i$ have to be introduced for every item, where $x_i = 1$ if item $i$ is in the

test, $x_i = 0$ if item $i$ is not in the test. $I$ is the number of items in the item bank. All test specifications also have to be expressed in terms of the $x_i$'s. For example, the variance functions, that is, the diagonal elements of the variance-covariance matrix, can be rewritten such that:

$$Var(\hat{\theta}_1|\boldsymbol{\theta}) = \frac{\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right)}{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2}, \quad (5)$$

and

$$Var(\hat{\theta}_2|\boldsymbol{\theta}) = \frac{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)}{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2}, \quad (6)$$

where the sums are taken over the items in the item bank.

**Objective Function**

As described above, for precise measurement of multiple traits, either Fisther's information matrix should be maximized or the variance-covariance matrix should be minimized. From optimum design theory, several criteria for optimality of matrices are known (see, for example, van der Linden, 1994). In this paper, A-optimality and D-optimality are considered.

The A-optimality criterion minimizes the trace of variance-covariance matrix, that is, the sum of the variance functions. By assigning weights to the variance functions, the relative importance of the different abilities can be specified. In this way, different cases of multidimensional test assembly can be treated. The optimallity criterion for the two-dimensional case can be formulated in the following manner:

$$\min w_1 Var(\theta_1|\boldsymbol{\theta}) + w_2 Var(\theta_2|\boldsymbol{\theta}). \quad (7)$$

When traits are considered equally important, both weights are set equal to each other. A different case occurs when items are sensitive to multiple abilities, but the test is developed to measure only one intentional ability, than the weights of the other abilities can be set equal to zero. These and other cases are described in van der Linden (1996). Besides these cases, weights can also be used in a different manner. When the magnitudes of both variance functions differ, weights can be used to rescale both functions in the criterion such that the larger term does not dominate the smaller one in the optimization process any more.

The function defined in Equations 7 is not only a function of $x_i$, but also a continuous function of the variables $(\theta_1, \theta_2)$. However, in the test assembly process it suffices to optimize these objective functions for a grid of points, instead of for the entire $\theta$-region. For example, Theunissen (1985) reduced the problem of maximizing the information function over the $\theta$-region the problem of maximizing the ability function at certain $\theta$-points. This technique can also be applied at a multidimensional $\boldsymbol{\theta}$-space. Let the two-dimensional grid be defined by $(s,t)$, where $s = 1, ..., S$ and $t = 1, ...T$. The resulting objective function is:

$$\min_{\substack{s=1..S \\ t=1..T}} w_1 Var(\theta_1|\boldsymbol{\theta}_{st}) + w_2 Var(\theta_2|\boldsymbol{\theta}_{st}) \tag{8}$$

This is a complicated objective function. Substituting both variance functions by Equation 5 and Equation 6 would result in:

$$\min \max_{\substack{s=1...S \\ t=1...T}} w_1 \frac{\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right)}{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2} + w_2 \frac{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)}{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2}. \tag{9}$$

It should be noted that the denominators of both terms are equal. So, the objective function can be rewritten into:

$$\min \max_{\substack{s=1..S \\ t=1..T}} \frac{w_1 \left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) + w_2 \left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)}{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2}. \quad (10)$$

The second criterion is the D-optimality. This criterion maximizes the determinant of Fishers information matrix, that is:

$$\max \det I(\boldsymbol{\theta}). \quad (11)$$

When this function is optimized for a grid of points $(s, t)$, the resulting objective function is:

$$\max \min_{\substack{s=1..S \\ t=1..T}} \det I(\boldsymbol{\theta}_{st}).$$

For the D-optimality no simplifications are needed. In the two-dimensional case, the objective function associated with D-optimality is equal to:

$$\max \min_{\substack{s=1..S \\ t=1..T}} \sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i \sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i - (\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i)^2. \quad (12)$$

**Model**

Several kinds of constraints can be formulated. In van der Linden (1998) categorical constraints, quantitative constraints and constraints on inter-item dependencies are distinguished. Categorical constraints deal with, for example, content classification, gender orientation, or minority orientation. For quantitative constraints one could think of response times or word count constraints. When items contain clues to each other they are in an enemy set and only one item from this set is allowed in a test.

Using the maximin approach (van der Linden & Boekkooi-Timminga, 1989), the following mathematical programming problem for A-optimality has to be solved:

$$\min \max_{\substack{s=1\ldots S \\ t=1\ldots T}} w_1 Var(\theta_1|\boldsymbol{\theta}_{st}) + w_2 Var(\theta_2|\boldsymbol{\theta}_{st}), \tag{13}$$

subject to:

$$\sum_{i\in C} x_i \leq n_C, \text{(categorical constraints)} \tag{14}$$

$$\sum_{i\in Q} q x_i \leq n_Q, \text{(quantitative constraints)} \tag{15}$$

$$\sum_{i\in E} x_i \leq 1, \text{(enemy sets)} \tag{16}$$

$$\sum_{i=1}^{I} x_i = n, \text{(test length)} \tag{17}$$

$$x_i \in \{0,1\}, \qquad i = 1\ldots I. \tag{18}$$

The parameters $n_C$ are the bounds that determine the number of items from the subset $C$ to be in the test. Bounds for the quantitative constraints are denoted by $n_Q$. Constraint 17 determines the test length and constraint 18 defines the decision variables. For the criterion of D-optimality a similar model can be described. However, the objective function is defined by Equation 12.

## Algorithms

In this paper, a new algorithm, based on linear approximation of the objective functions, is proposed. This algorithm can be applied at unconstrained test assembly problems, and at constrained ones. In order to evaluate the algorithm, two algorithm were used as benchmarks. The first algorithm is a generalization of Lueght's (Lueght,1996) greedy algorithm for MAT to the P&P case. The second algorithm randomly select

items from the item bank. Therefore, it is only applicable at unconstrained test assembly problems.

**Linear Approximation of the Objective Function**

The algorithm is based on linear programming techniques. These techniques are often applied in automated test assembly. However, they optimize a linear objective function subject to a number of constraints on the test attributes. In order to apply these techniques to the problem at hand, a linear approximation of the objective function in Equation 13 should be made. The general formula for the linear approximation of a function $f$ at a given point $\overline{\mathbf{x}}$ is given by $f(\overline{\mathbf{x}}) + \nabla f(\overline{\mathbf{x}})^t(\mathbf{x} - \overline{\mathbf{x}})$, where $\nabla$ is the vector of first order derivatives (See, for example, Bazaraa, Sheraldi and Shetty,.1993, page 121).

When a linear approximation of the objective function is calculated, a point $\boldsymbol{\theta}$ has to be chosen where the function is approximated. Unfortunately it is unclear in advance, which $\boldsymbol{\theta}$-point to choose. Therefore, it was decided to optimize the worst performance of the linear approximation of the objective functions over the gridpoints $(s, t)$. The linear approximation of the simplified objective functions in Equation 10 and Equation 12 are given in Appendix A.

In Appendix A it is shown that the approximations only differ in the values of $k_{jst}$. So, the resulting test assembly problem for both the objective functions in Equation 10 and Equation 12 can be formulated as follows:

$$\min \max_{s,t} k_{1st} \sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i + k_{2st} \sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i + k_{3st} \sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i \qquad (19)$$

subject to:

$$\sum_{i \in C} x_i \leq n_C, \text{(categorical constraints)} \qquad (20)$$

$$\sum_{i \in Q} q x_i \leq n_Q, \text{(quantitative constraints)} \qquad (21)$$

$$\sum_{i \in E} x_i \leq 1, \text{(enemy sets)} \qquad (22)$$

$$\sum_{i=1}^{I} x_i = n, \text{(test length)} \tag{23}$$

$$x_i \in \{0,1\}, \qquad i = 1 \ldots I. \tag{24}$$

where the coefficients $k_{jst}$ are determined by the linear approximation (see Appendix A). Now the objective function is a linear function of the decision variables $x_i$, and general automated test assembly techniques can be used to solve the problem.

**Greedy algorithm**

For the assembly of adaptive tests that measure multiple abilities, different approaches are at hand. In Segall (1996), a locally optimal item selection procedure for MAT is described. Each time an item is selected that provides the largest decrement in the volume of the credibility ellipsoid. In van der Linden (1999) the item is selected that minimizes the variance of a linear combination of the abilities. In Luecht (1996), Segall's approach is extended to the constrained case, by building the total test content constraints into the objective function. In these procedures the item is selected that contributes most to the objective function in each iteration. Therefore they are in the class of the so-called greedy algorithms.

The adaptive strategies can be applied rather straightforwardly in the context of assembling tests. Items that contribute most to the objective function have to be sequentially selected, until the maximum number of items in the test is reached. In case of no constraints this greedy heuristic selects the item whose value of

$$\max_{\substack{s=1\ldots S \\ t=1\ldots T}} \frac{w_1 \left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) + w_2 \left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)}{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2}$$

is minimal. Or when the objective function in Equation 12 is used, the heuristic selects the item whose value of

$$\min_{\substack{s=1..S \\ t=1..T}} \sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i \sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2$$

is maximal. In case of constraints a different strategy should be used. All constraints can be built into the objective function (Luecht, 1996). The composite objective function is optimized. When this strategy is applied to the problem in equation 13 to 18, the following problem has to be solved:

$$\min \sum_{j=1}^{4} \alpha_j d_j$$

subject to:

$$\begin{aligned}
T - \min \max_{\substack{s=1\ldots S \\ t=1\ldots T}} w_1 Var(\theta_1|\boldsymbol{\theta}_{st}) + w_2 Var(\theta_2|\boldsymbol{\theta}_{st}) &= d_1, \\
\max(\sum_{i \in C} x_i - n_C, 0) &= d_2, \\
\max(\sum_{i \in Q} q_i x_i - n_Q, 0) &= d_3, \\
\max(\sum_{i \in E} x_i - 1, 0) &= d_4, \\
\sum_{i=1}^{I} x_i &= n, \\
x_i &\in \{0, 1\},
\end{aligned}$$

where $T$ is a prespecified target for the objective function, $d_j$ is the deviation of the $j$-th constraint, and $\alpha_j$ denotes the weight of the deviation of the $j$-th constraint. Like the non-constraint case, each iteration the item is added to the test that minimizes the objective function. A target $T$ for the objective function can be obtained by solving the non-constraint problem first.

**Random Item Selection Algorithm**

Items are randomly selected from the item bank until the maximum number of items is reached. Since it is impossible to take constraints into account in this algorithm, it is only applicable at unconstrained test assembly problems. The tests resulting from this algorithm can be used as benchmarks. In order to be usefull, a proposed algorithm should

produce test that are at least as good as the tests provided by the random item selection algorithm.

## Empirical Examples

An ACT Assessment Program Mathematics Item Pool was used to assemble tests from. The item pool consisted of 176 items. The calibration of this items was carried out with the program NOHARM (Fraser and McDonald, 1988), and an acceptable fit was obtained by a two dimensional version of the model in Equation 1. The items were classified according to content specification and skill.

First, the problem without constraints was solved for the A-criterion and the D-criterion. The linear approximation, the greedy heuristic and the random selection algorithm were applied. In this way the loss due to the linear approximation of the objective function was examined for both optimality criteria. Second, the effects of adding constraints to the model were investigated for the greedy heuristic and the linear approximation. Therefore several sets of constraints were added to the problem (See Table 1).

The main program for solving these problems was written in PASCAL 7.0 and heuristic seven of the program Contest (Timminga, van der Linden and Schweizer, 1996) was used to solve the linear programming parts of the examples.

### Example 1

In the first exampel tests were assembled over the complete grid of points defined by $(\theta_1, \theta_2) \in \{-1, 0, 1\} \times \{-1, 0, 1\}$. The test had to contain a fixed number of items. No further constraints on item or test attributes were defined. The number of items in the test varied from 10 to 50. The results of a heuristic that randomly selected the fixed number of items from the item pool were added to serve as a bench mark. In Figure 1 the resulting values for the A-criterion objective function are shown.

==========

Insert Figure 1 at about here

==========

As can be seen, solving the test assembly problem with a linear approximation of the A-criterion objective function does not provide good tests. The approximation performed hardly better than random item selection algorithm. The greedy algorithm provided much better results. For the D-criterion the results are shown in Figure 2. Remind that the objective of the second criterion is to maximize the objective function instead of to minimize, so the higher the results, the better.

==========

Insert Figure 2 at about here

==========

For this criterion, the linear approximation performed much better. The assembled tests contained much more information than a random selection of items from the pool, and almost as much information as the greedy test.

**Example 2**

For the D-criterion both the greedy algorithm and the linear approximation were compared for a number of additional constraints. The following content and skill constraints were added to the problem:

(1) The test should contain at least $n_{PG}$ plane geometry, $n_{PA}$ pre-algebra, $n_{EA}$ elementary algebra, $n_{CG}$ coordinate geometry, $n_{TG}$ trigonometry, and $n_{IA}$ intermediate algebra items.

(2) At least $n_{BS}$ basic skill items, $n_{AP}$ application items, and $n_{AN}$ analysis items should be included in the test.

Hence the following additional constraints were obtained:

$$\sum_{i \in V_{PG}} x_i \geq n_{PG}, \tag{25}$$

$$\sum_{i \in V_{PA}} x_i \geq n_{PA}, \tag{26}$$

$$\sum_{i \in V_{EA}} x_i \geq n_{EA}, \tag{27}$$

$$\sum_{i \in V_{CG}} x_i \geq n_{CG}, \tag{28}$$

$$\sum_{i \in V_{TG}} x_i \geq n_{TG}, \tag{29}$$

$$\sum_{i \in V_{IA}} x_i \geq n_{IA}, \tag{30}$$

$$\sum_{i \in V_{BA}} x_i \geq n_{BS}, \tag{31}$$

$$\sum_{i \in V_{AP}} x_i \geq n_{AP}, \tag{32}$$

$$\sum_{i \in V_{AN}} x_i \geq n_{AN}, \tag{33}$$

where for example $V_{PG}$ is the set indices of the items with content classification Plane Geometry (PG). Several sets of constraints were tested. Each violation of a constraint was counted as a fault. The weights of all constraints in the greedy heuristic were set equal to one. The lowest value of the D-criterion on the grid of $\theta$-points defined by $(-1, 0, 1) \times (-1, 0, 1)$ and the number of faults are presented in Table 1. In this example the test length was set equal to twenty-five.

==========

Insert Table 1 at about here.

==========

The greedy heuristic assembled tests with a higher value of the D-criterion. When the constraints were easy to meet, that is, when enough items were present in the item pool, no violations of constraints were made. However, when the constraints were hard to meet, violations occured when the greedy heuristic was used. For the sixth set of constraint even seven faults were counted.

## Discussion

In this paper, a new algorithm for assembling tests from item pools that are calibrated using a multidimensional IRT model was proposed . The performance of the algorithm was compared with a greedy algorithm for both the A-optimality criterion and the D-optimality criterion. In the first example both algorithms were compared without any constraints in the model. For A-optimality the linear approximation did not prove to be useful. For D-optimality good tests were obtained.

An explanation of these performances can be found in the formulas of both criteria. When the formulas in Equation 10 and Equation 12 are compared, it is easy to see that the first formula is far more difficult to approximate by a linear function than the second formula. Therefore, the linear approximation performed much worse for the A-criterion than for the D-criterion. The conclusion can be drawn that the algorithm based on a linear approximation of the objective function only performs well for simple nonlinear functions. In both examples the items were calibrated with a two-dimensional model. Since higher-dimensional models result in more difficult formulas, the question how this heuristic works out for higher-dimensional models needs additional research.

When the results of the second example are compared with the unconstrained case, the differences between both algorithms are increased. In the unconstrained case, the values for D-optimality were 2.13 for the greedy algorithm and 1.83 for the linear approximation. The difference is 0.30. In the constrained cases the differences vary from 0.67 to 0.95. An explanation can be found in the way the tests are assembled by both algorithms. Because the greedy algorithm allows violations of constraints, it is less restricted by the constraints than the linear approximation. Therefore, the difference increases. As a result of this, it is hard to compare the algorithms.

It depends on the item pool and the preferences of the test assembler which of both algorithms should be applied in practise. In large scale testing programs, different versions of a test should be comparable and it is important that all constraints are met. When the item pool is well designed and contains enough items to fulfill all constraints the greedy algorithm should be applied. The first two sets of constraints in the second

example illustrate this case. No violations were made and the resulting tests were far more informative than the linear approximation. On the other hand, when many constraints are formulated for the test and the item pool is hardly able to fulfill them, the greedy algorithm will result in many faults. For these cases the algorithm based on the linear approximation of the objective function should be applied. What remains are the cases where the greedy algorithm results in a few faults. For these cases no general recommendations can be given and the preferences of the test assembler will be decisive.

## Appendix A.

For a test the objective function for the A-optimality criterion is stated in Equation 10. It is formulated in the following manner:

$$\min \max_{\substack{s=1..S \\ t=1..T}} \frac{w_1 \left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) + w_2 \left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)}{\left(\sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i\right)\left(\sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i\right) - \left(\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i\right)^2},$$

where the sums are taken over the items in the test. Three terms are present in this funcion. Define:

$$\begin{aligned} x &\equiv \sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i, \\ y &\equiv \sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i, \\ z &\equiv \sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i. \end{aligned}$$

For a given ability point $(s,t)$, and a given test, the functions $x, y$, and $z$ can be calculated. The result is denoted by $(\overline{x}, \overline{y}, \overline{z})$.

The objective function can be rewritten into:

$$\min \max_{\substack{s=1..S \\ t=1..T}} f(x,y,z)$$

where

$$f(x,y,z) \equiv \frac{w_1 y + w_2 x}{xy - z^2}.$$

The linear approximation of the objective function in the point $(\overline{x}, \overline{y}, \overline{z})$ is equal to:

$$\min \max_{\substack{s=1..S \\ t=1..T}} \frac{\partial f}{\partial x}(\overline{x}, \overline{y}, \overline{z}) \cdot x + \frac{\partial f}{\partial y}(\overline{x}, \overline{y}, \overline{z}) \cdot y + \frac{\partial f}{\partial z}(\overline{x}, \overline{y}, \overline{z}) \cdot z + c$$

where $c \equiv f(\overline{x},\overline{y},\overline{z}) + \nabla f(\overline{x},\overline{y},\overline{z})^t \cdot (\overline{x},\overline{y},\overline{z})$, and the partial derivatives are given by:

$$\begin{aligned}
k_{1st} &= \frac{\partial f}{\partial x}(\overline{x}_{st},\overline{y}_{st},\overline{z}_{st}) = \frac{w_2}{\overline{x}_{st}\cdot\overline{y}_{st}-\overline{z}_{st}^2} - \frac{w_2\cdot\overline{x}_{st}\cdot\overline{y}_{st}}{(\overline{x}_{st}\cdot\overline{y}_{st}-\overline{z}_{st}^2)^2} \\
k_{2st} &= \frac{\partial f}{\partial y}(\overline{x}_{st},\overline{y}_{st},\overline{z}_{st}) = \frac{w_1}{\overline{x}_{st}\cdot\overline{y}_{st}-\overline{z}_{st}^2} - \frac{w_1\cdot\overline{x}_{st}\cdot\overline{y}_{st}}{(\overline{x}_{st}\cdot\overline{y}_{st}-\overline{z}_{st}^2)^2} \\
k_{3st} &= \frac{\partial f}{\partial z}(\overline{x}_{st},\overline{y}_{st},\overline{z}_{st}) = \frac{2\cdot\overline{z}_{st}\cdot(w_1\overline{x}_{st}+w_2\overline{y}_{st})}{(\overline{x}_{st}\cdot\overline{y}_{st}-\overline{z}_{st}^2)^2}
\end{aligned}$$

Tests resulting from the greedy algorithm can be used to calculate $k_{1st}$, $k_{2st}$, and $k_{3st}$.

For D-optimality the objective function is stated in Equation 12;

$$\max \min_{\substack{s=1..S\\t=1..T}} \sum_{i=1}^{I} a_{2i}^2 P_i Q_i x_i \sum_{i=1}^{I} a_{1i}^2 P_i Q_i x_i - (\sum_{i=1}^{I} a_{1i} a_{2i} P_i Q_i x_i)^2.$$

In terms of $x$, $y$ and $z$ the function is equal to:

$$\max \min_{\substack{s=1..S\\t=1..T}} xy - z^2.$$

The partial derivatives, that define the linear approximation, are given by:

$$\begin{aligned}
k_{1st} &= \frac{\partial f}{\partial x}(\overline{x}_{st},\overline{y}_{st},\overline{z}_{st}) = \overline{y}_{st} \\
k_{2st} &= \frac{\partial f}{\partial y}(\overline{x}_{st},\overline{y}_{st},\overline{z}_{st}) = \overline{x}_{st} \\
k_{3st} &= \frac{\partial f}{\partial z}(\overline{x}_{st},\overline{y}_{st},\overline{z}_{st}) = -2\cdot\overline{z}_{st}
\end{aligned}$$

The coefficients $k_{jst}$ are equal to the partial derivatives $\nabla f$ evaluated at the points $(s,t)$ for a given reference test.

## References

Ackerman, T.A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement*, 18, 257-275.

Bazaraa, M.S., Sherali, H.D., and Shetty, C.M. (1979). *Nonlinear programming: Theory and Algorithms.* New York, NY: Wiley.

Birnbaum, A. *Some latent trait models and their use in inferring an examinee's ability.* In F.M. Lord & M.R. Novick, Statistical theories of mental test scores (pp. 397-479). Reading MA: Addison-Wesley.

Bock, R.D., Gibbons, R.and Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.

Fraser, C. and McDonald, R.P. (1988). *NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: University of New England, Centre for Behavioral Studies.

Lehmann, E.L. (1983). *Theory of point estimation.* New York NY: Wiley.

Lord, F.M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Lueght, R.M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389-404.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.

McKinley, R.L. & Reckase, M.N. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space.* (Research Rep. ONR 83-2). Iowa City IA: American College Testing.

Reckase, M.D. (1979). Unifactor latent trait models applied to multi-factortests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.

Reckase, M.D. (1985). The difficulty of test items that measure more than one dimension *Applied Psychological Measurement*, 9, 401-412.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*,52, 589-617.

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika*, 50, 411-420.

Timminga, E. , van der Linden, W.J., adn Schweizer, D.A. (1996). *ConTEST 2.0: A decision support system for item banking and optimal test assembly (computer program and manual)*. Groningen, The Netherlands: iec ProGAMMA.

van der Linden, W.J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, 20, 373-388.

van der Linden, W.J. (1998). Optimal Assembly of Psychological an dEducational Tests. *Applied Psychological Measurement*, 22, 195-211.

van der Linden, W.J. (1999). Muldidimensional adaptive testing with a minimum error variance criterium. *Journal of Educational and Behavioral Statistics*, 24, (in press).

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., and Tissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilson, D., Wood, R., and Gibbons, R. (1987). *TESTFACT*. Mooresville IN: Scientific Software.

Yen, W.M. (1983). Use of the three-parameter model in the development of standardized achievement test. In R.K. Hambleton (Ed.), *Applications of item response theory*, (pp. 123-141). Vancouver, BC: Educational Research Institute of British Columbia.

Table 1.

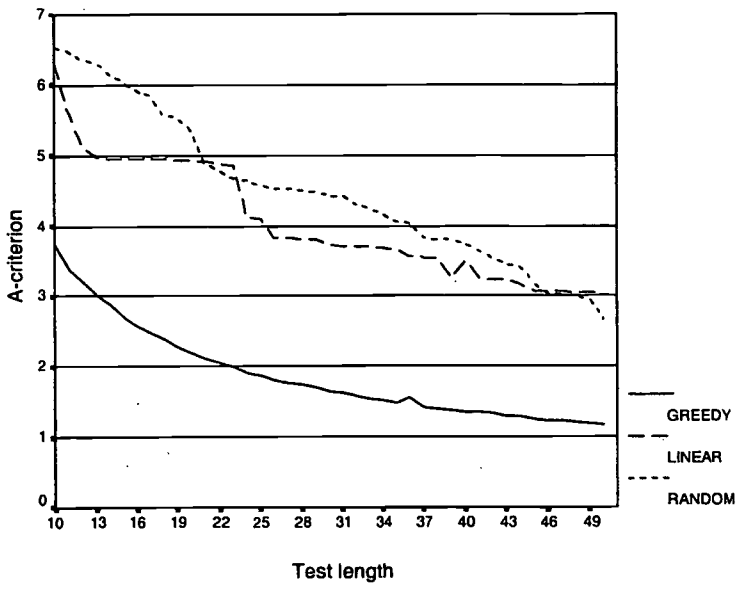
Results of both algorithms for different sets of constraints.

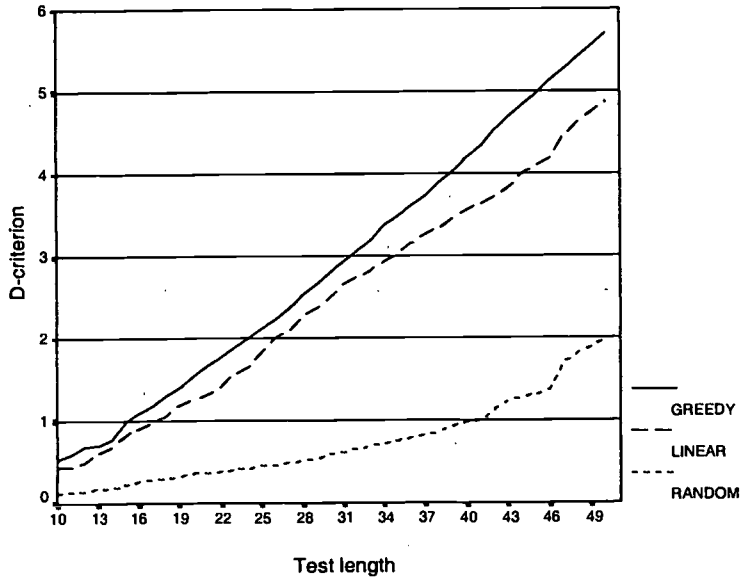
| Sets of constraints | | | | | | | | | Greedy | | Linear | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PG | PA | EA | CG | TG | IA | BS | AP | AN | det(I) | # Faults | det(I) | # Faults |
| 2 | 2 | 2 | 2 | 2 | 2 | 7 | 7 | 2 | 2.13 | - | 1.18 | - |
| 3 | 3 | 3 | 3 | 3 | 3 | 9 | 9 | 5 | 1.83 | - | 1.19 | - |
| 4 | 4 | 4 | 4 | 4 | 4 | 10 | 10 | 5 | 1.91 | 3 | 1.18 | - |
| 9 | 2 | 2 | 2 | 2 | 8 | 15 | 7 | 3 | 1.80 | 1 | 1.21 | - |
| 4 | 4 | 4 | 4 | 4 | 4 | 8 | 7 | 10 | 1.75 | 5 | 0.94 | - |
| 3 | 3 | 3 | 3 | 10 | 3 | 8 | 8 | 8 | 1.51 | 7 | 0.76 | - |
| 1 | 1 | 10 | 1 | 11 | 1 | 16 | 8 | 1 | 1.43 | 3 | 0.76 | - |

## Figure Captions

*Figure 1:* Results of different algorithms for the A-criterion.

*Figure 2* : Results of different algorithms for the D-criterion.





**Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.**

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

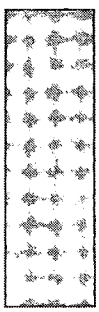
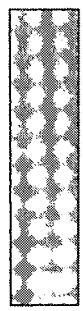
RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

*faculty of*

# EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

**U.S. Department of Education**
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM032319

# NOTICE

## REPRODUCTION BASIS

This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").